THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB Article

What Makes a Functional Gene Regulatory Network? A Circuit Motif Analysis

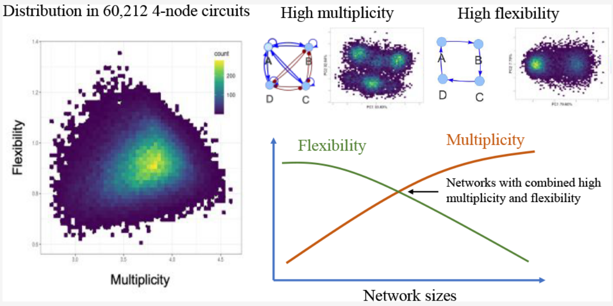
Published as part of The Journal of Physical Chemistry virtual special issue "Jose Onuchic Festschrift".

Lijia Huang, Benjamin Clauss, and Mingyang Lu*

Cite This: *J. Phys. Chem. B* 2022, 126, 10374−10383 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: One of the key questions in systems biology is to understand the roles of gene regulatory circuits in determining cellular states and their functions. In previous studies, some researchers have inferred large gene networks from genome wide genomics/transcriptomics data using the top-down approach, while others have modeled core gene circuits of small sizes using the bottom-up approach. Despite many existing systems biology studies, there is still no general rule on what sizes of gene networks and what types of circuit motifs a system would need to achieve robust biological functions. Here, we adopt a gene circuit motif analysis to discover four-node circuits responsible for multiplicity (rich in dynamical behavior), flexibility (versatile to alter gene expression), or both. We identify the most reoccurring two-node circuit motifs and the co-occurring motif pairs. Furthermore, we investigate the contributing factors of multiplicity and flexibility for large gene networks of different types and sizes. We find that gene networks of intermediate sizes tend to have combined high levels of multiplicity and flexibility. Our study will contribute to a better understanding of the dynamical mechanisms of gene regulatory circuits and provide insights into rational designs of robust gene circuits in synthetic and systems biology.



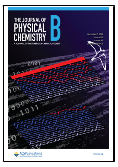
INTRODUCTION

It has been established that many important biological cellular processes are controlled by complex gene regulatory networks (GRNs).[1,2] The inference of the GRNs driving cellular state transitions has become one of the major challenges in systems biology.[3] To address this question, many bioinformatics methods[4−12] have been developed to infer GRNs using genomics data sets, such as gene expression data. Yet, researchers have found it very difficult to generate high-quality network models.[13−15] In our view, the main issue is that most bioinformatics methods rely on statistical tests to determine whether one gene regulates another but seldom evaluate whether an inferred GRN can operate as a functional dynamical system. This view is supported by a recent benchmark study of GRN inference methods in that current existing methods often perform poorly to recover ground-true networks.[13]

The above-mentioned concern has led us to think about what properties of a GRN would contribute to a functional system. There are two features of a GRN worth looking into. First, a functional GRN needs to generate rich dynamical behaviors, e.g., multiple steady states (i.e., multistability) and/or oscillatory states. As shown in earlier studies, random GRNs tend to generate less interesting dynamical behaviors than biological networks.[16,17] On the other hand, multistability is often required for a GRN model to capture a variety of cellular states during cell differentiation.[16,18−21] Second, a functional GRN needs to be sufficiently flexible so that the GRN can be controlled by extrinsic cell signaling or an environmental factor.[22−25] It is quite common that the activation of a signaling pathway can drive the transition of cellular states.[26] Equally typical are gene knockdown/knockout experiments designed to understand the functions of genes based on the effects of gene perturbations.[27] Thus, functional GRNs need to be flexible, even in the presence of a certain level of compensation and adaptation due to network redundancy.[28,29] Therefore, it is reasonable to hypothesize that a functional GRN is required to produce rich dynamics and meanwhile be flexible upon perturbations.

Here, under this conceptual framework, we adopted our recently developed gene circuit motif analysis approach[30] to explore nonredundant four-node gene circuits that are responsible for multiplicity (i.e., being rich in dynamical behavior), flexibility (i.e., being versatile to alter gene

Received: July 30, 2022
Revised: November 21, 2022
Published: December 5, 2022



https://doi.org/10.1021/acs.jpcb.2c05412

expression), or both. There are many previous studies on circuit motif analysis;[2,18,31,32] however, we here focused on the properties of multiplicity and flexibility by extensive simulations and statistical analysis. From the identified small circuits, we will determine the most reoccurring two-node circuit motifs and the propensity of co-occurrence of two circuit motifs. Furthermore, using the identified circuit motifs, we generated a variety of large GRNs of different types (linear, scale-free, and random) and different sizes, from which we investigated the contributing factors of the multiplicity and flexibility of large GRNs. We hope that the outcomes of these analyses will shed light on the improved modeling of biological GRNs.

■ METHODS

A Quantitative Circuit Motif Analysis. We have recently developed a new approach for gene circuit motif analysis,[30] which allows identifying reoccurring two-node circuit motifs and patterns of motif coupling from the ranking of 60212 nonredundant four-node gene circuits by a certain dynamical feature. In this approach, we numerate all possible nonredundant four-node gene circuits, and for each circuit, we generate the steady-state gene expression profiles for an ensemble of 10000 mathematical models with the random circuit perturbation (RACIPE) method[33,34] (see the section RACIPE Simulations and SI Text 1 for details). On the basis of a user-defined scoring function computed from the simulated gene expression data (see the section Defining Network Multiplicity and Flexibility for the two scores defined in this study), we can rank all the four-node circuits and identify two-node circuit motifs enriched in the top-ranking circuits. A similar enrichment analysis can also be applied to evaluate the co-occurrence of two circuit motifs. Our approach has several advantages over some existing methods. First, to ensure a robust statistical analysis, the circuit motif analysis utilizes extensive simulation data from all nonredundant four-node gene circuits. Second, the ensemble-based circuit simulations allow us to quantify circuits' dynamical behavior not specific to a special set of kinetic parameters. A scoring function defined in this way enables us to rank gene circuits robustly and efficiently. Third, from the analysis of all four-node gene circuits, we can evaluate the enrichment of small circuit motifs and their coupling. Note that we limit our analysis to four-node gene circuits without any signaling node (a node without a regulator) or target node (a node without a target gene), as such a circuit could be reduced to a circuit of a smaller size. Therefore, the circuit motifs we would explore here could be complementary but also distinct from the most significant circuit motifs identified in the previous studies.[2] In this study, we applied this enrichment analysis to identify small circuit motifs contributing to a functional regulatory system.

RACIPE Simulations. For any four-node gene regulatory circuit or a large gene regulatory network (GRN), we computed its gene expression distribution by using *ra*ndom *ci*rcuit *pe*rturbation (RACIPE)[33,34] (sRACIPE 1.12). RACIPE simulates an ensemble of mathematical models for a gene circuit/network with randomly selected kinetic parameters and obtains steady-state gene expression profiles. Compared to the traditional mathematical modeling approaches, RACIPE simulated gene expression profiles are derived from the same ODEs but with different kinetic parameters to capture extrinsic factors, such as cell-to-cell variations and different environmental conditions. We have previously shown, in a few examples of biological networks, that the gene expression profiles derived from RACIPE simulations form distinct clusters of gene expression patterns, where the cluster can be associated with experimentally observed cellular states of the systems.[16,33,35−37] Thus, RACIPE is a convenient and powerful method to evaluate the behavior of gene circuits/networks from simulated gene expression distributions. Here, we generated 10000 gene expression profiles (log-transformed and standardized) for each circuit/network to compute its gene expression distribution. A summary of the RACIPE implementation, including the ODEs, the ranges of model parameters, and the choice of initial conditions, is presented in SI Text 1.

Defining Network Multiplicity and Flexibility. For each gene circuit, we applied RACIPE to generate the steady-state gene expression profiles of 10000 mathematical models with randomly generated kinetic parameters. As mentioned above, RACIPE-simulated gene expression profiles from a biological network usually form robust clusters of gene expression patterns. However, the gene expression profiles from random gene networks are usually less structured.[16,17] To reflect the ability of biological GRNs in generating distinct cellular states, we defined a scoring function H, namely multiplicity, by the negative differential entropy[38] of the simulated gene expression distributions of the 10000 models:

$$H = \langle \log(p_i) \rangle \sim \frac{1}{N} \sum_i \log\left(\frac{k}{NVR(k)^d} \right) \tag{1}$$

where p_i is the local density of model i, $\langle\ \rangle$ denotes average, N is the number of models, and the summation is over all simulated models. Here, the local density p_i is computed by the k nearest neighbors ($k$nn) estimator,[39] where $R(k)$ is the Euclidean distance of gene expression profiles between kth nearest model and the center model, d is the dimension of the gene expression space ($d = 4$ for any four-node gene circuit), and a constant scaling factor $V = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$. The multiplicity defined in eq 1 can be interpreted by the mean log local density. The higher the overall local densities, the higher the H values. Moreover, in the situation of high local density, more gene expression clusters can be observed. This is consistent with our previous findings that the local densities of the gene expression profiles simulated from a stem cell gene regulatory circuit are overall larger than those from a random gene circuit.[16]

We next defined the flexibility of a gene circuit, F, by the extent of changes in the gene expression distributions of 10000 RACIPE models between the unperturbed and knockdown (KD) conditions. More specifically, the flexibility F is defined as

$$F = \sum_{j=1}^{d} \sum_{l=1}^{d} e_l D(p_{l,0}, p_{l,j}) \tag{2}$$

where the summations are over all gene nodes j (from 1 to the dimension d) and all principal components (PCs) l (from 1 to d). Here, principal component analysis is performed on the gene expression data of models from the unperturbed condition. e_l is the lth eigenvalue, which we incorporated here to emphasize the changes along the largest PCs. We quantified the differences in gene expression distributions by D, the Kolmogorov−Smirnov test[40] of the probability

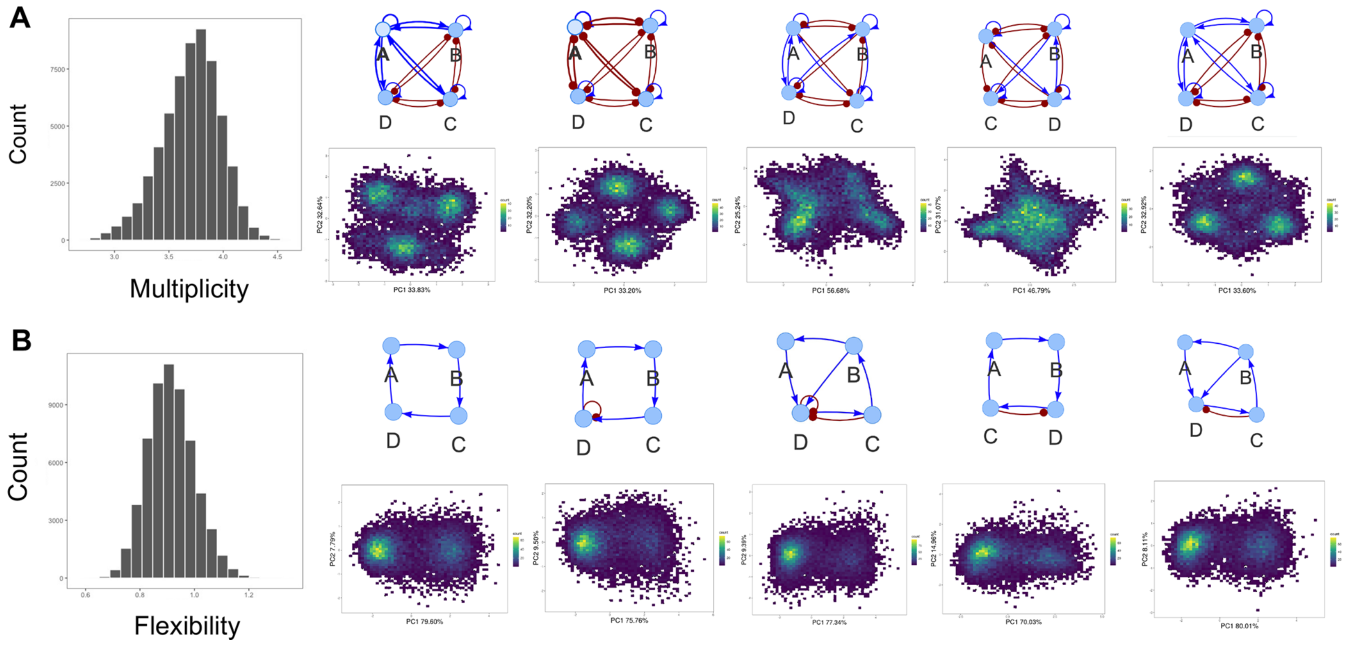


Figure 1. Multiplicity and flexibility of gene regulatory circuits. (A) The leftmost plot shows the histogram of multiplicity for all nonredundant four-node circuits. The right panels show, for the top five circuits ranked by multiplicity, the circuit diagrams (top row) and the density maps of RACIPE-simulated gene expression projected onto the first two PCs (bottom row). In the circuit diagrams, the nodes represent genes, labeled as A, B, C, and D. The blue lines and arrows represent excitatory regulations, and the red lines and dots represent inhibitory regulations. Panel B shows the outcomes for the flexibility score.

distribution of the data along each PC between the unperturbed condition ($p_{l,0}$ for the lth PC) and the perturbed condition, in which gene j is knocked down ($p_{l,j}$). Here, we subset 10% models with the lowest production rates of the KD gene j to compute the distribution for the KD condition.

In addition, we also defined another scoring function for the combined multiplicity and flexibility. Because the circuits' multiplicity H and flexibility F have values in different ranges, we chose to rank circuits with a new score G, defined by the product of H and F:

$$G = HF \tag{3}$$

Enrichment Analysis of Circuit Motifs. To explore gene circuit motifs associated with circuit multiplicity or flexibility, we performed an extensive circuit motifs enrichment analysis on all 60212 nonredundant four-node gene circuits. These circuits exclude those that can be equivalently reduced to circuits of three or a smaller number of nodes (see SI Text 2 and the previous study[30] for more details). For each circuit, we performed RACIPE simulations to generate gene expression profiles and evaluated the H, F, and G scores defined by eqs 1−3.

For each of the above scores (denoted below as Q), we calculated the occurrence of any two-node circuit motifs (listed in Figure S1) within the top-ranked four-node circuits by the score. There are several existing dedicated methods for network motif detection.[41] But we chose to utilize a simple numeration approach based on adjacency matrices because we focused on very small circuits in this study, and our analysis can account for different edge types (i.e., activation and inhibition) and autoregulations. We computed the enrichment score for each two-node circuit motif, defined as

$$E = \log\left(\frac{\sum_l 1 - H^{-}(Q, Q_0, n)}{\sum_l H^{-}(Q, Q_0, n)}\right) \tag{4}$$

where $H^{-}(x, x_0, n) := 1/[1 + (x/x_0)^n]$ is the inhibitory Hill function. The Hill threshold, Q_0, is set to be the Q value of the 600th ranked circuit. The Hill coefficient n is set to 20, for a sharp transition near the threshold Q_0. Moreover, we applied the enrichment analysis to determine the enriched co-occurrence of two two-node circuit motifs.

When performing the enrichment analysis, we evaluated statistical significance by a permutation test as follows. First, enrichment scores were calculated for every two-node motif as described above. Second, a null distribution was created by shuffling the ranking indices of all four-node circuits. Third, enrichment scores were calculated again but with the shuffled indices. Fourth, steps two and three were repeated for 10000 times. Fifth, p values were computed by the fraction of cases where the enrichment score of each motif from the shuffled indices is greater than the original enrichment score with the unshuffled indices. Adjusted p values were then computed for multiple hypothesis testing by the BH method.[42] A similar strategy for statistical significance test has been utilized in earlier studies.[43] Details of the enrichment analysis are also described in our previous study.[30]

Generating Large GRNs. We programmatically generated three types of large GRNs: random, scale-free, and sequential networks, where the procedure is illustrated in Figure S2. To generate the random networks, we first built *skeleton* networks using standard network generation algorithms, and then each node in a skeleton network is replaced with a gene circuit motif of choice. For an edge connecting two circuit motifs, a randomly selected gene from the first circuit motif was linked to a randomly selected gene from the second circuit motif.

https://doi.org/10.1021/acs.jpcb.2c05412
J. Phys. Chem. B 2022, 126, 10374−10383

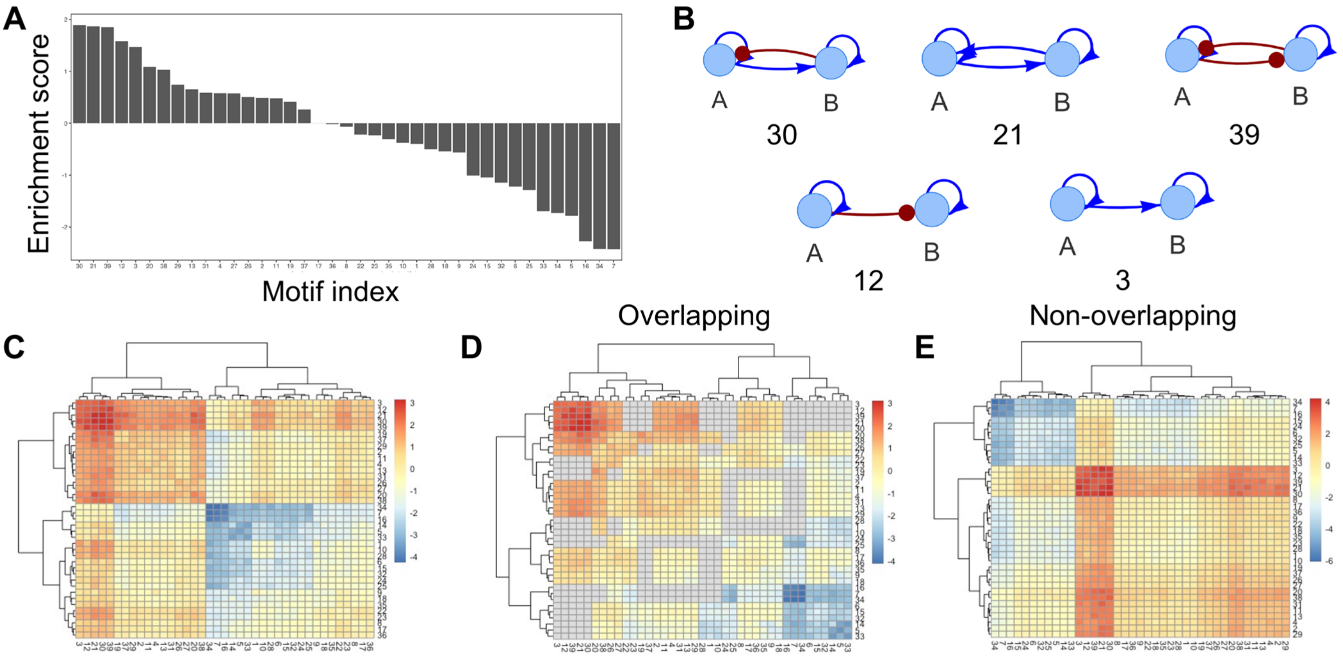


Figure 2. Circuit motif enrichment analysis with respect to circuit multiplicity. (A) Enrichment scores for all two-node circuit motifs were computed from all nonredundant four-node gene circuits ranked by multiplicity. The enrichment is significant for most motifs (adjusted p values <0.01, as shown in Figure S6A), except for motifs #17 and 36 (circuit diagrams and indices in Figure S1). (B) Diagrams of the top five enriched circuit motifs. (C−E) Heatmaps of the enrichment scores for the co-occurrence of all pairs of two two-node circuit motifs. Three heatmaps correspond to the overall co-occurrence (C), the co-occurrence of two motifs with a shared node (overlapping, D), and the co-occurrence of two motifs without a shared node (nonoverlapping, E). Hierarchical clustering analysis was applied to each case with the Euclidean distance and complete linkage. There are gray colors in panel D, as some motif combinations do not exist.

First, to generate the skeleton networks with Gaussian degree distribution and directed edges, we used the *erdos.renyi.game* function from *igraph* R package[44] and the *gnp* or *gnm* method. For the *gnm* version of the random networks (denoted as *random ver1*), the total number of edges was set to equal the total number of nodes. Therefore, these GRNs are sparsely connected. For the *gmp* version of the random networks (denoted as *random ver2*), the probability of an edge occurring between nodes was set to 30%. Therefore, these GRNs are densely connected. From a skeleton network, half of the edges were randomly selected and designated as inhibitory edges, while the other half as excitatory edges. Second, to generate the scale-free networks, we used the *sample_pa* function from the *igraph* R package to generate the skeleton networks with the power law degree distribution and directed edges. Third, the sequential networks were constructed by connecting the desired number of two-node circuit motifs one after another. An edge (either excitatory or inhibitory) was added to connect a randomly picked gene from the first motif (source) to a randomly picked gene from the second motif (target). Afterward, another edge (either excitatory or inhibitory) was added to connect the other gene from the second motif (source) to a randomly picked gene from the third motif (target). We continued the procedure iteratively to connect all motifs.

Altogether, we generated networks of four types (random ver1, random ver2, scale-free, and sequential), five different network sizes (10, 20, 30, 40, and 50 nodes), and with three sets of circuit motifs as the building blocks. The first set of motifs contains the top three enriched motifs by multiplicity, the second set contains the top three enriched motifs by flexibility, and the third set contains all the above six motifs (see the Results section for details of the motif enrichment analysis). Each kind of network was generated randomly for ten times; thus, we analyzed on a total of 600 large GRNs (4 × 5 × 3 × 10).

RESULTS

Characterizing Circuit Multiplicity. We applied the multiplicity scoring function to all 60212 nonredundant four-node gene circuits. As shown in Figure 1A, the multiplicity score H ranges from about 2.5 to 4.8, and the distribution of multiplicity is a negatively skewed unimodal distribution, with slightly more circuits of high H values. We found that the scoring function H is indeed effective in capturing circuit multiplicity (Figure S3). We observed that the topmost circuits ranked by H tend to contain regulatory links of mutual activations, mutual inhibitions, and self-activations. The tight regulatory connectivity in those circuits allows them to have a higher number of gene expression clusters, as shown in the density map of the simulated gene expression profiles projected onto the first two principal components (PCs) (Figure 1A, right panels). Instead, the bottommost circuits ranked by H tend to have only one gene expression cluster (Figure S4A).

Next, we applied the circuit motif enrichment analysis by comparing the occurrence of any two-node circuit motifs within the topmost four-node circuits ranked by multiplicity with the occurrence within the rest circuits. As shown in Figure 2A,B (significant test in Figure S6A), the top five enriched two-node circuit motifs all contain self-activations, suggesting its dominant role in determining high multiplicity. The top three motifs (#30, 21, and 39; see Figure S1 for the circuit diagrams and indices of all two-node motifs), which have similarly high enrichment scores, all contain mutual regulatory links.

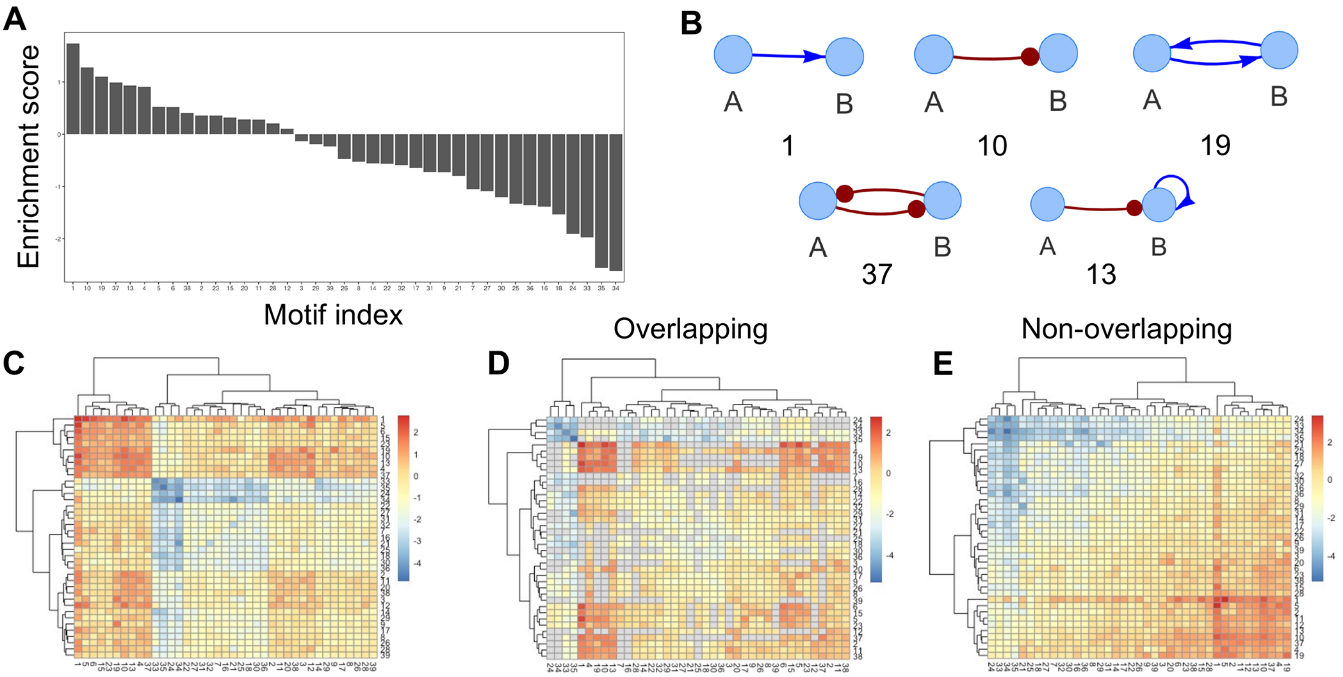


Figure 3. Circuit motif enrichment analysis with respect to circuit flexibility. Contents similar to those in Figure 2. Here, all nonredundant four-node gene circuits are scored and ranked by flexibility. The enrichment of a single motif is significant for all motifs (adjusted p values <0.01, as shown in Figure S6B).

Contrarily, the bottom three motifs (#7, 34, and 16) all contain self-inhibitions. These findings are all consistent with what we observed in the top-ranked four-node circuits in Figure 1A and with previous studies showing that self-activation generates multistability and self-inhibition stabilizes the gene expression state.[2,30,45] Furthermore, we evaluated the enrichment of the co-occurrence of two circuit motifs within the top-ranked four-node circuits by the multiplicity, as shown in Figure 2C−E. We observed again that two motifs with self-activation tend to be more enriched, while two motifs with self-inhibition tend to be less enriched.

Characterizing Circuit Flexibility. We applied the flexibility scoring function to rank all 60212 nonredundant four-node gene circuits. As shown in Figure 1B, the flexibility score F ranges from about 0.6 to 1.4, and the distribution of flexibility is close to a symmetric unimodal distribution. We tested the flexibility score F on a few four-node circuits (Figure S5) and found that for circuits with larger F, gene expression distributions have noticeably larger changes upon gene KD perturbations. We observed that the topmost circuits ranked by F tend to be sparsely connected. Compared to the topmost circuits ranked by H, the topmost circuits by F usually have a monodirectional interaction between two nodes (either the first node regulating the second or the second node regulating the first), and there are fewer autoregulations. Interestingly, the circuits with high flexibility usually have gene expression profiles of two clusters (Figure 1B, right panels), while the circuits with low flexibility tend to have gene expression profiles of multiple clusters (Figure S4B). This observation can be understood as follows. For circuits allowing only one gene expression cluster, the possible gene expression distribution is limited by the cluster. For circuits with a higher number of gene expression clusters, it is hard to transit through multiple states by perturbation. Thus, circuits with two gene expression clusters are the most likely to achieve substantial changes in gene expression distributions upon perturbations.

Next, we applied the circuit motif enrichment analysis to circuits ranked by flexibility. As shown in Figure 3A,B (significant test in Figure S6B), the first and second most enriched circuit motifs (#1 and 10) all contain a single regulatory link from one gene to the other. The third and fourth most enriched circuit motifs (#19 and 37) are circuits with mutual activation and mutual inhibition (toggle switch), respectively. We noticed slight differences in the enrichment scores between circuits with excitatory regulations and those with inhibitory regulations, presumably because of sampling deviations and the measurement of flexibility by knockdown perturbations. These most enriched circuit motifs usually do not prefer autoregulation. Interestingly, the toggle-switch-like circuit motifs (#19 and 37) are frequently observed in the topmost flexible circuits, as they are known to generate bistability.[46,47] These motifs ensure circuits to be sparsely connected and bistable, thus allowing the whole circuit to be also flexible. Furthermore, we evaluated the enrichment of the co-occurrence of two circuit motifs within the top-ranked four-node circuits by the flexibility, as shown in Figure 3C−E. Interestingly, we observed frequent co-occurrence of motif #1 with three motifs, #5, 6, and 1 itself. These three circuit motifs all share the same excitatory regulatory link from one gene to the other but differ by just a self-inhibitory link. The co-occurrence of these motifs further demonstrates the sparseness of regulatory interactions as one of the determining factors of flexible circuits.

Distinct Circuits with Features of Combined Multiplicity and Flexibility. In the previous two sections, we have evaluated the multiplicity and flexibility of all nonredundant four-node gene circuits and applied motif enrichment analysis to identify two-node circuit motifs associated with either high multiplicity or high flexibility. Furthermore, we evaluated the

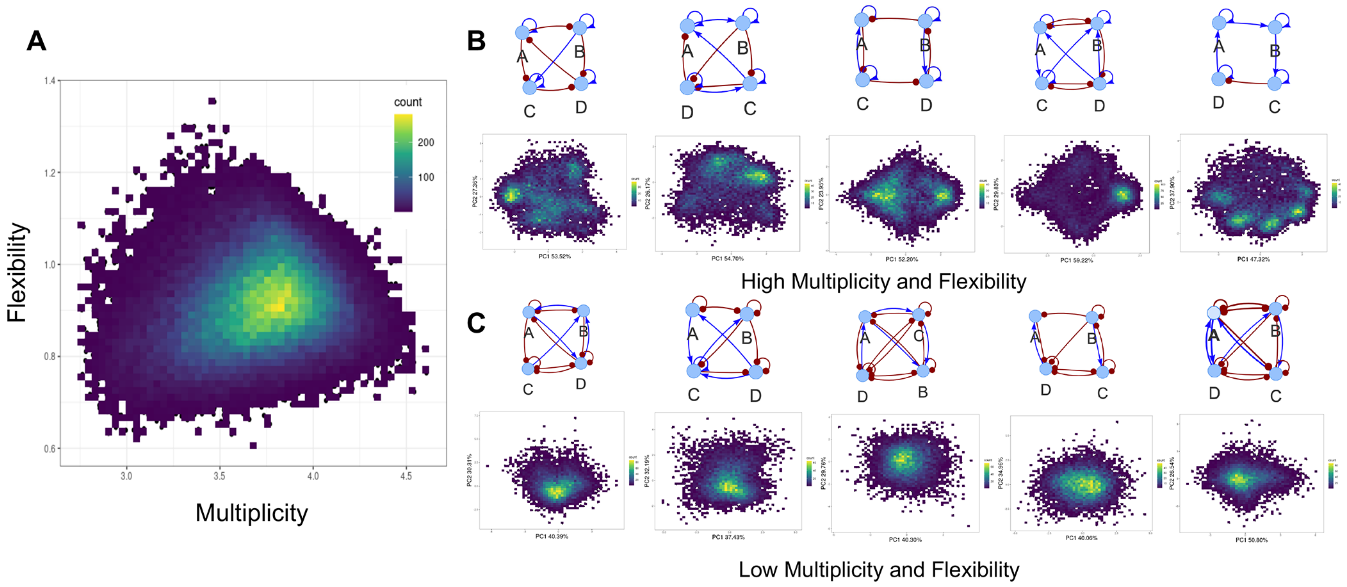


Figure 4. Gene regulatory circuits ranked by combined multiplicity and flexibility. (A) The density map of multiplicity (x-axis) and flexibility (y-axis) for all nonredundant four-node gene circuits. (B) The panels show, for the top five circuits ranked by the product of multiplicity and flexibility, the circuit diagrams (top row) and the density maps of RACIPE-simulated gene expression projected onto the first two PCs (bottom row). Panel C shows the outcomes for the bottom five circuits.

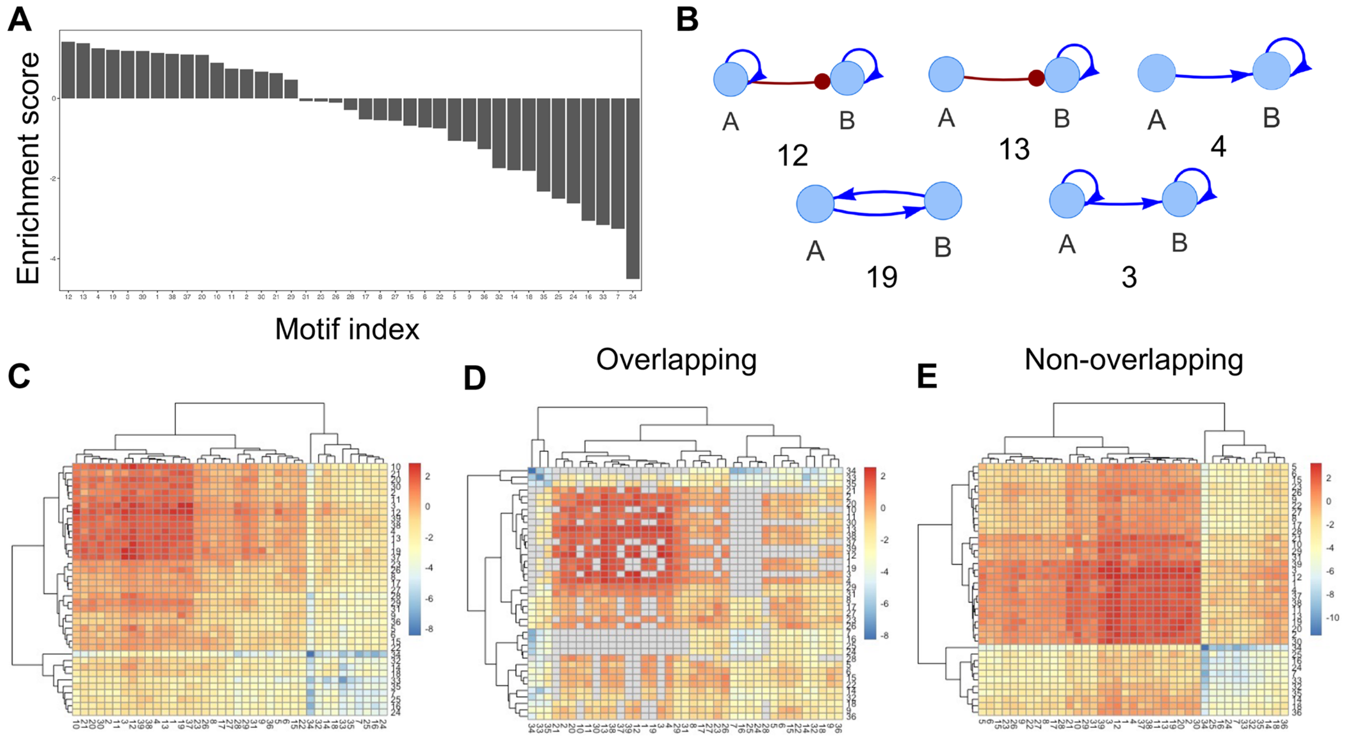


Figure 5. Circuit motif enrichment analysis by both multiplicity and flexibility. Contents similar to those in Figure 2. Here, all nonredundant four-node gene circuits are scored and ranked by the product of multiplicity and flexibility. The enrichment of a single motif is significant for most motifs, except for motif #31 (adjusted p values <0.01, as shown in Figure S6C).

relationship between the multiplicity and flexibility of a circuit. From the density and scatter plot in Figure 4A, we observed a weak correlation (0.17275 Pearson correlation coefficient) between these two scores. Interestingly, we found circuits rarely have high multiplicity and flexibility simultaneously (only 0.048% of circuits with both H and F higher than 1.5 standard deviations above the mean value). However, much more circuits were found to have high multiplicity and low flexibility (0.144% of circuits with H higher than 1.5 standard deviations above the mean value and F lower than 1.5 standard deviations below the mean value). More circuits were also found to have low multiplicity and high flexibility (0.332% circuits with H lower than 1.5 standard deviations below the mean value and F higher than 1.5 standard deviations above the mean value). It is reasonable that despite no apparent correlation between multiplicity and flexibility, circuits with the

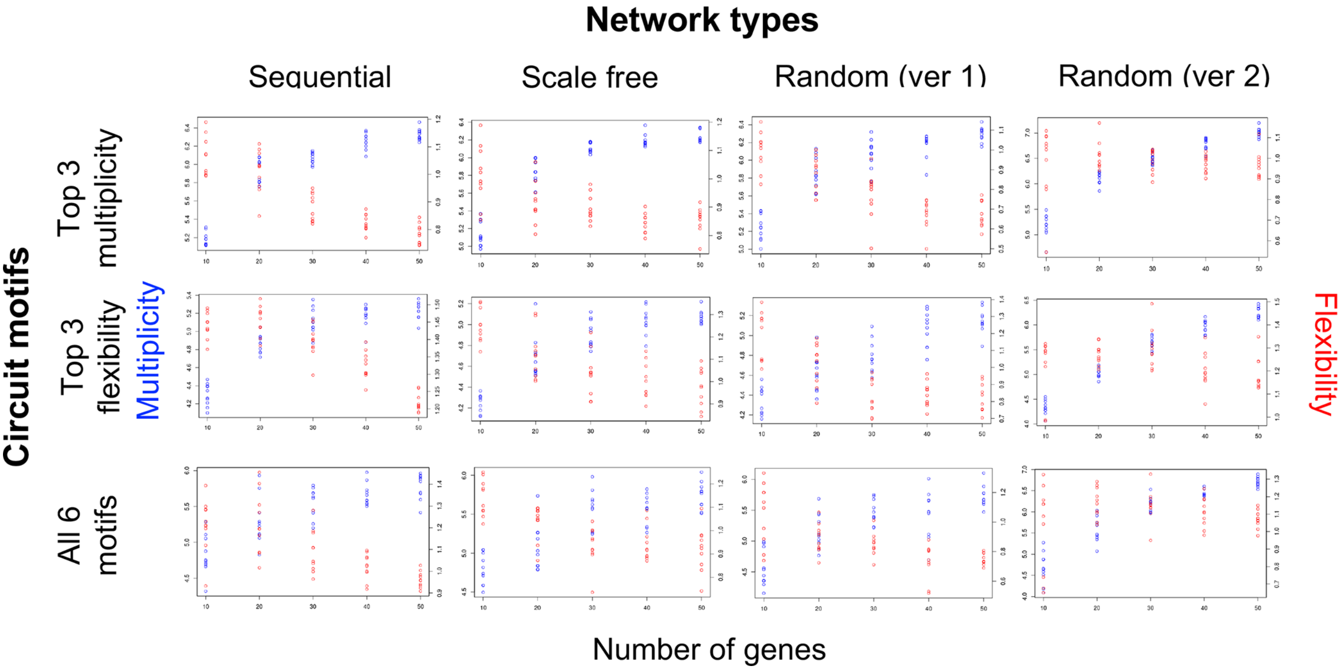


Figure 6. Multiplicity and flexibility of large gene regulatory networks. The plots show the multiplicity (blue points) and flexibility (red points) for GRNs of different sizes (number of genes on the *x*-axis) and different network types (each panel). Different columns correspond to sequential networks (1st column), scale-free networks (2nd column), random networks with sparse connectivity (ver 1, 3rd column), and random networks with dense connectivity (ver 2, 4th column). Different rows correspond to networks synthesized with top 3 multiplicity two-node motifs (1st row), top 3 flexibility two-node motifs (2nd row), and all these 6 motifs (3rd row).

highest multiplicity are less likely to be flexible, while circuits with the highest flexibility are required to have fewer gene expression clusters, thus being low in multiplicity.

As we discussed earlier, we are interested in functional GRNs with both high multiplicity and flexibility, despite their low occurrence in four-node circuits. We applied the circuit motif analysis with the combined score G defined in eq 3. The range of G is from about 1.5 to 6.72. The top five ranked circuits, as shown in Figure 4B, have the following features. First, these circuits all have relatively simpler circuit topologies compared to the topmost circuits ranked by multiplicity. Second, these circuits all contain multiple self-activations, thus generating a high number of gene expression clusters. Third, the gene expression distributions resulting from these circuits seem to be less structured compared to those from the circuits with the highest multiplicity. All these features contribute to being high in both multiplicity and flexibility. Contrarily, the bottom five ranked circuits, as shown in Figure 4C, have more self-inhibitions, allow for gene expression distributions of a single cluster, and have highly connected circuit topologies. These properties are exactly opposite to those of top-ranked circuits, explaining why the bottom-ranked circuits have low multiplicity and flexibility. These observations are also consistent with the outcomes of the circuit motif enrichment analysis, as shown in Figure 5 (significant test in Figure S6C). Note that we identified circuit motifs #19 and 39 again as enriched motifs with high multiplicity and flexibility. These toggle-switch-like motifs were observed presumably because they can generate bistability, thus having more potential to generate more states when coupled with similar motifs, and meanwhile can allow flexible switches among states.[21,48]

Multiplicity and Flexibility in Large Random GRNs. Lastly, we explored the properties of multiplicity and flexibility in large random GRNs. To generate GRNs of an extended range of multiplicity and flexibility, we selected a list of two-node circuit motifs as the building blocks and synthesized them into large GRNs of different sizes, with either sequential, scale-free, or random topological structure (see the Methods section for the detailed implementation). The selected motifs are either (1) the top 3 two-node circuit motifs ranked by multiplicity (i.e., motifs # 30, 21, and 39), (2) the top 3 two-node circuit motifs ranked by flexibility (i.e., motifs #1, 10, and 19), or (3) the motifs from both (1) and (2). For each motif type, GRN topology type, and GRN size, we randomly generated the topology of ten networks (see the companion GitHub repository[49] for all GRN topologies), followed by RACIPE simulations to generate 10000 gene expression profiles for each GRN. To calculate the multiplicity and flexibility for a large GRN, we performed principal component analysis on the standardized log transformation gene expression and applied eqs 1 and 2 using the data projected onto the first four PCs. Note that in eq 2 the summation over gene perturbation is still applied to all genes in the GRN. We chose to first project data onto the first four PCs, as the data with reduced dimensions usually capture gene expression states well. Moreover, low-dimensional reduction has been widely used in high-dimensional gene expression data analysis.

The multiplicity and flexibility of these random GRNs are summarized in Figure 6, where we identified the following interesting findings. First, the overall trends of multiplicity and flexibility for different GRN sizes and types are very similar for different choices of circuit motifs. The multiplicity scores are usually at high levels when using the motifs with the highest multiplicity and at low levels when using the motifs with the highest flexibility. Similarly, the flexibility scores are usually at high levels when using the motifs with the highest flexibility and at low levels when using the motifs with the highest multiplicity. Thus, among GRNs of different sizes, multiplicity and flexibility are anticorrelated. Our finding also suggests that the multiplicity/flexibility properties of a large GRN are largely determined by the properties of the circuit motifs within the GRN.

Second, regardless of the type of GRNs, multiplicity was found to be linearly correlated to the number of genes but saturated for large number of genes (blue points in Figure 6). For each category of GRNs (i.e., different sizes and motif types), the variations of multiplicity among ten random networks are mostly small, but slightly larger for the GRNs with mixed motifs. We also computed the multiplicity when the local density was estimated with the gene expression profiles of all dimensions (Figure S7), and in this case, multiplicity is always linearly correlated to the number network genes. The dependence of multiplicity on GRN sizes can be understood as follows. When the GRNs are very small, the number of distinct states allowed by the GRNs is also limited. When the GRNs become larger, much richer network behaviors can be observed, therefore larger multiplicity. However, when the GRNs get extremely large, although the variations of gene expression still increase (multiplicity for data with full dimensions), the number of distinct gene expression states gets saturated (multiplicity for data with reduced dimensions).

Third, flexibility was found to be linearly anticorrelated to the number genes for sequential networks, scale-free networks, and random networks where motifs are sparsely connected with a fixed number of interactions per motif denoted as *random ver1*; see the Methods section or details) (red points, first to third columns in Figure 6), despite much larger variations in flexibility among ten networks of the same category. In those situations, we also observed a saturation of flexibility for small GRNs. Because of high variations and saturation of flexibility, we also observed a few small GRNs with low flexibility. Interestingly, for networks where motifs are densely connected with a fixed ratio of interactions per motif (denoted as *random ver2*), we observed a bell shape of flexibility with respect to the number of genes; i.e., the highest flexibility may occur in GRNs of intermediate sizes.

Taken together, when the network size increases, multiplicity increases while flexibility decreases. Both multiplicity and flexibility tend to be saturated for large and small GRNs, respectively. On the basis of these findings, we perceive that the GRNs with both high multiplicity and flexibility are likely of intermediate sizes.

■ DISCUSSION

In this study, we explored the types of gene circuit motifs that contribute to a functional gene regulatory network (GRNs). We first defined two scoring functions to quantify the multiplicity and flexibility of a gene regulatory circuit based on the circuit's gene expression distribution. We then systematically applied the scores to rank all nonredundant four-node gene circuits. By applying gene circuit motif analysis, we identified reoccurring two-node circuit motifs and the co-occurrence of two motifs that enriched in top-ranked circuits by either multiplicity, flexibility, or a combination of both. Furthermore, using the enriched motifs as the building blocks, we generated many GRNs of different types and sizes and investigated the GRN properties that contribute to high levels of multiplicity and flexibility. We hope this study will improve our understanding of the design of biological GRNs.

The core approach utilized in this study is the circuit motif enrichment analysis that we recently introduced.[30] We have demonstrated the effectiveness of this approach in identifying not only circuit motifs associated with a particular dynamical behavior but also the coupling of two circuit motifs. Here, we focused on multiplicity, the ability of a GRN in generating a high number of states, and flexibility, the ability of a GRN in altering gene expression upon perturbations. In our view, multiplicity and flexibility are among the most important features of a functional GRN. From the enrichment analysis, circuit motifs with mutual regulations and self-activation tend to have high multiplicity, while circuit motifs with single monodirectional regulation and without autoregulation tend to have high flexibility. Remarkably, two types of circuit motifs allow both high multiplicity and high flexibility—either motifs with sparse connectivity and self-activation or toggle-switch-like motifs.

While it is important to elucidate the types of circuit motifs having high multiplicity and/or flexibility, we also wonder how these circuit motifs contribute to the multiplicity and flexibility of larger GRNs. To address this question, we generated GRNs of different sizes and types using the enriched circuit motifs as the building blocks. From an extensive network analysis, we found that network multiplicity and flexibility indeed are largely impacted by the types of circuit motifs with the GRNs. Overall, GRNs of intermediate sizes (around 30; also see Figure 6) tend to have combined high levels of multiplicity and flexibility. Thus, we hypothesize that a biological GRN, when considered as a functional dynamical system, should be of intermediate sizes. This can be understood by the following: when a GRN is too small, it is not complex enough to robustly generate desired functionality; when a GRN is too large, it could be too rigid to allow sufficient control by external signals or environmental factors.[50,51] Thus, GRNs of intermediate sizes can alleviate the issues of smaller and larger GRNs. In our view, this criterion of network size would be helpful to elucidate the design principle of biological GRNs and improve the effectiveness of GRN inference.

There are a few related topics that are worth further investigation. First, when simulating circuit dynamics, we assume AND logics when multiple genes regulate a target gene. It is interesting to evaluate how other types of logical rules[52] affect GRN multiplicity and flexibility. Second, the current approach focuses on characterizing gene expression distributions, but many functional GRNs may act as oscillators.[53−55] One of the potential future directions is to evaluate oscillatory dynamics[56] in the circuit motif analysis. Third, we have observed that multiplicity get saturated for large networks. Indeed, biological networks usually exhibit a limited number of cellular states, thus limiting the level of multiplicity. It is worth some further studies to elucidate the saturation of cellular states in biological networks.[17]

■ ASSOCIATED CONTENT

Data Availability Statement

Relevant R code from this study is available at the following link: https://github.com/huanglijiaU201614513/circuitanalysis. That includes the code for RACIPE simulations, state distribution scoring, construction of all four-node circuit motifs, enrichment analysis of nonredundant four-node circuits, and generation of random gene networks. The topologies of all large gene networks are also provided.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.2c05412.

> SI Text 1: details of modeling gene circuits using RACIPE; SI Text 2: generation of all nonredundant four-node gene circuits; Figure S1: circuit diagrams and indices of all two-node motifs; Figure S2: schematic of random network generation; Figure S3: five four-node gene circuits of different multiplicity scores; Figure S4: nonredundant four-node circuits with the least multiplicity and flexibility ranks; Figure S5: four four-node gene circuits of different flexibility scores; Figure S6: adjusted *p* values for the enrichment of all two-node circuit motifs; Figure S7: multiplicity and flexibility of large gene regulatory networks (where multiplicity scores were computed using all dimensions) (PDF)

AUTHOR INFORMATION

Corresponding Author

Mingyang Lu − *Center for Theoretical Biological Physics and Department of Bioengineering, Northeastern University, Boston, Massachusetts 02115, United States; Genetics Program, Graduate School of Biomedical Sciences, Tufts University, Boston, Massachusetts 02111, United States; The Jackson Laboratory, Bar Harbor, Maine 04609, United States*; orcid.org/0000-0001-8158-0593; Phone: (617) 373-8017; Email: m.lu@northeastern.edu

Authors

Lijia Huang − *Center for Theoretical Biological Physics and Department of Bioengineering, Northeastern University, Boston, Massachusetts 02115, United States*

Benjamin Clauss − *Center for Theoretical Biological Physics, Northeastern University, Boston, Massachusetts 02115, United States; Genetics Program, Graduate School of Biomedical Sciences, Tufts University, Boston, Massachusetts 02111, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.2c05412

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award R35GM128717 and by startup funds from Northeastern University.

REFERENCES

(1) Chuang, H.-Y.; Hofree, M.; Ideker, T. A Decade of Systems Biology. *Annu. Rev. Cell Dev Biol.* **2010**, *26*, 721−744.

(2) Shen-Orr, S. S.; Milo, R.; Mangan, S.; Alon, U. Network Motifs in the Transcriptional Regulation Network of Escherichia Coli. *Nat. Genet.* **2002**, *31* (1), 64−68.

(3) Karlebach, G.; Shamir, R. Modelling and Analysis of Gene Regulatory Networks. *Nat. Rev. Mol. Cell Biol.* **2008**, *9* (10), 770−780.

(4) Aibar, S.; González-Blas, C. B.; Moerman, T.; Huynh-Thu, V. A.; Imrichova, H.; Hulselmans, G.; Rambow, F.; Marine, J.-C.; Geurts, P.; Aerts, J.; et al. SCENIC: Single-Cell Regulatory Network Inference and Clustering. *Nat. Methods* **2017**, *14* (11), 1083−1086.

(5) Ding, H.; Douglass, E. F.; Sonabend, A. M.; Mela, A.; Bose, S.; Gonzalez, C.; Canoll, P. D.; Sims, P. A.; Alvarez, M. J.; Califano, A. Quantitative Assessment of Protein Activity in Orphan Tissues and Single Cells Using the MetaVIPER Algorithm. *Nat. Commun.* **2018**, *9* (1), 1471.

(6) Ding, J.; Aronow, B. J.; Kaminski, N.; Kitzmiller, J.; Whitsett, J. A.; Bar-Joseph, Z. Reconstructing Differentiation Networks and Their Regulation from Time Series Single-Cell Expression Data. *Genome Res.* **2018**, *28* (3), 383−395.

(7) Chan, T. E.; Stumpf, M. P. H.; Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst* **2017**, *5* (3), 251−267.

(8) Huynh-Thu, V. A.; Irrthum, A.; Wehenkel, L.; Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS One* **2010**, *5* (9), e12776.

(9) Kim, S. Ppcor: An R Package for a Fast Calculation to Semi-Partial Correlation Coefficients. *Commun. Stat Appl. Methods* **2015**, *22* (6), 665−674.

(10) Matsumoto, H.; Kiryu, H.; Furusawa, C.; Ko, M. S. H.; Ko, S. B. H.; Gouda, N.; Hayashi, T.; Nikaido, I. SCODE: An Efficient Regulatory Network Inference Algorithm from Single-Cell RNA-Seq during Differentiation. *Bioinformatics* **2017**, *33* (15), 2314−2321.

(11) Specht, A. T.; Li, J. LEAP: Constructing Gene Co-Expression Networks for Single-Cell RNA-Sequencing Data Using Pseudotime Ordering. *Bioinformatics* **2016**, *33* (5), 764−766.

(12) Qiu, X.; Rahimzamani, A.; Wang, L.; Ren, B.; Mao, Q.; Durham, T.; McFaline-Figueroa, J. L.; Saunders, L.; Trapnell, C.; Kannan, S. Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. *Cell Syst* **2020**, *10* (3), 265−274.

(13) Pratapa, A.; Jalihal, A. P.; Law, J. N.; Bharadwaj, A.; Murali, T. M. Benchmarking Algorithms for Gene Regulatory Network Inference from Single-Cell Transcriptomic Data. *Nat. Methods* **2020**, *17* (2), 147−154.

(14) Kang, Y.; Thieffry, D.; Cantini, L. Evaluating the Reproducibility of Single-Cell Gene Regulatory Network Inference Algorithms. *Frontiers in Genetics* **2021**.

(15) Seçilmiş, D.; Hillerton, T.; Sonnhammer, E. L. L. GRNbenchmark - a Web Server for Benchmarking Directed Gene Regulatory Network Inference Methods. *Nucleic Acids Res.* **2022**, *50* (W1), W398−W404.

(16) Huang, B.; Lu, M.; Galbraith, M.; Levine, H.; Onuchic, J. N.; Jia, D. Decoding the Mechanisms Underlying Cell-Fate Decision-Making during Stem Cell Differentiation by Random Circuit Perturbation. *Journal of The Royal Society Interface* **2020**, *17* (169), 20200500.

(17) Tripathi, S.; Kessler, D. A.; Levine, H. Biological Networks Regulating Cell Fate Choice Are Minimally Frustrated. *Phys. Rev. Lett.* **2020**, *125* (8), 088101.

(18) Ye, Y.; Kang, X.; Bailey, J.; Li, C.; Hong, T. An Enriched Network Motif Family Regulates Multistep Cell Fate Transitions with Restricted Reversibility. *PLoS Comput. Biol.* **2019**, *15* (3), e1006855.

(19) Duddu, A. S.; Sahoo, S.; Hati, S.; Jhunjhunwala, S.; Jolly, M. K. Multi-Stability in Cellular Differentiation Enabled by a Network of Three Mutually Repressing Master Regulators. *J. R Soc. Interface* **2020**, *17* (170), 20200631.

(20) Laurent, M.; Kellershohn, N. Multistability: A Major Means of Differentiation and Evolution in Biological Systems. *Trends Biochem. Sci.* **1999**, *24* (11), 418−422.

(21) Guantes, R.; Poyatos, J. F. Multistable Decision Switches for Flexible Control of Epigenetic Differentiation. *PLOS Computational Biology* **2008**, *4* (11), e1000235.

(22) Bhalla, U. S.; Iyengar, R. Emergent Properties of Networks of Biological Signaling Pathways. *Science* **1999**, *283* (5400), 381−387.

(23) Li, M.; Gao, H.; Wang, J.; Wu, F.-X. Control Principles for Complex Biological Networks. *Brief Bioinform* **2019**, *20* (6), 2253−2266.

(24) Zañudo, J. G. T.; Yang, G.; Albert, R. Structure-Based Control of Complex Networks with Nonlinear Dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (28), 7234−7239.

(25) Liu, Y.-Y.; Barabási, A.-L. Control Principles of Complex Systems. *Rev. Mod. Phys.* **2016**, *88* (3), 035006.

(26) Zhang, W.; Liu, H. T. MAPK Signal Pathways in the Regulation of Cell Proliferation in Mammalian Cells. *Cell Res.* **2002**, *12* (1), 9−18.

(27) Adli, M. The CRISPR Tool Kit for Genome Editing and Beyond. *Nat. Commun.* **2018**, *9* (1), 1911.

(28) Bao, Y.; Hayashida, M.; Liu, P.; Ishitsuka, M.; Nacher, J. C.; Akutsu, T. Analysis of Critical and Redundant Vertices in Controlling Directed Complex Networks Using Feedback Vertex Sets. *J. Comput. Biol.* **2018**, *25* (10), 1071−1090.

(29) Bhattacharya, P.; Raman, K.; Tangirala, A. K. Discovering Adaptation-Capable Biological Network Structures Using Control-Theoretic Approaches. *PLOS Computational Biology* **2022**, *18* (1), e1009769.

(30) Clauss, B.; Lu, M.A Quantitative Evaluation of Topological Motifs and Their Coupling in Gene Circuit State Distributions. bioRxiv July 20, 2022; p 2022.07.19.500691.

(31) Schaerli, Y.; Munteanu, A.; Gili, M.; Cotterell, J.; Sharpe, J.; Isalan, M. A Unified Design Space of Synthetic Stripe-Forming Networks. *Nat. Commun.* **2014**, *5* (1), 4905.

(32) Jiménez, A.; Cotterell, J.; Munteanu, A.; Sharpe, J. A Spectrum of Modularity in Multi-Functional Gene Circuits. *Molecular Systems Biology* **2017**, *13* (4), 925.

(33) Huang, B.; Jia, D.; Feng, J.; Levine, H.; Onuchic, J. N.; Lu, M. RACIPE: A Computational Tool for Modeling Gene Regulatory Circuits Using Randomization. *BMC Syst. Biol.* **2018**, *12* (1), 74.

(34) Kohar, V.; Lu, M. Role of Noise and Parametric Variation in the Dynamics of Gene Regulatory Circuits. *NPJ. Syst. Biol. Appl.* **2018**, *4*, 40.

(35) Katebi, A.; Kohar, V.; Lu, M. Random Parametric Perturbations of Gene Regulatory Circuit Uncover State Transitions in Cell Cycle. *iScience* **2020**, *23* (6), 101150.

(36) Ramirez, D.; Kohar, V.; Lu, M. Toward Modeling Context-Specific EMT Regulatory Networks Using Temporal Single Cell RNA-Seq Data. *Front Mol. Biosci* **2020**, *7*, 54.

(37) Su, K.; Katebi, A.; Kohar, V.; Clauss, B.; Gordin, D.; Qin, Z. S.; Karuturi, R. K. M.; Li, S.; Lu, M.NetAct: A Computational Platform to Construct Core Transcription Factor Regulatory Networks Using Gene Activity. bioRxiv May 9, 2022; p 2022.05.06.487898.

(38) Lord, W. M.; Sun, J.; Bollt, E. M. Geometric K-Nearest Neighbor Estimation of Entropy and Mutual Information. *Chaos* **2018**, *28* (3), 033114.

(39) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *American Statistician* **1992**, *46* (3), 175−185.

(40) Dodge, Y.*The Concise Encyclopedia of Statistics*; Springer Science & Business Media: 2008.

(41) Yu, S.; Feng, Y.; Zhang, D.; Bedru, H. D.; Xu, B.; Xia, F. Motif Discovery in Networks: A Survey. *Computer Science Review* **2020**, *37*, 100267.

(42) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. 57. *Journal of the Royal Statistical Society Series B* **1995**, *57*, 289−300.

(43) Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (43), 15545−15550.

(44) Gabor Csardi, T. N. The Igraph Software Package for Complex Network Research. *InterJournal* **2006**, 1695.

(45) Alon, U. Network Motifs: Theory and Experimental Approaches. *Nat. Rev. Genet* **2007**, *8* (6), 450−461.

(46) Huang, B.; Lu, M.; Jia, D.; Ben-Jacob, E.; Levine, H.; Onuchic, J. N. Interrogating the Topological Robustness of Gene Regulatory Circuits by Randomization. *PLoS Comput. Biol.* **2017**, *13* (3), e1005456.

(47) Gardner, T. S.; Cantor, C. R.; Collins, J. J. Construction of a Genetic Toggle Switch in Escherichia Coli. *Nature* **2000**, *403* (6767), 339−342.

(48) Tian, X.-J.; Zhang, X.-P.; Liu, F.; Wang, W. Interlinking Positive and Negative Feedback Loops Creates a Tunable Motif in Gene Regulatory Networks. *Phys. Rev. E* **2009**, *80* (1), 011926.

(49) *Github repository of this study*. https://github.com/huanglijiaU201614513/circuitanalysis.

(50) Cooke, J.; Nowak, M. A.; Boerlijst, M.; Maynard-Smith, J. Evolutionary Origins and Maintenance of Redundant Gene Expression during Metazoan Development. *Trends Genet* **1997**, *13* (9), 360−364.

(51) Kafri, R.; Levy, M.; Pilpel, Y. The Regulatory Utilization of Genetic Redundancy through Responsive Backup Circuits. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (31), 11653−11658.

(52) Wang, N.; Lefaudeux, D.; Mazumder, A.; Li, J. J.; Hoffmann, A. Identifying the Combinatorial Control of Signal-Dependent Transcription Factors. *PLOS Computational Biology* **2021**, *17* (6), e1009095.

(53) Li, Z.; Yang, Q. Systems and Synthetic Biology Approaches in Understanding Biological Oscillators. *Quant Biol.* **2018**, *6* (1), 1−14.

(54) Ferrell, J. E.; Tsai, T. Y.-C.; Yang, Q. Modeling the Cell Cycle: Why Do Certain Circuits Oscillate? *Cell* **2011**, *144* (6), 874−885.

(55) Bell-Pedersen, D.; Cassone, V. M.; Earnest, D. J.; Golden, S. S.; Hardin, P. E.; Thomas, T. L.; Zoran, M. J. Circadian Rhythms from Multiple Oscillators: Lessons from Diverse Organisms. *Nat. Rev. Genet* **2005**, *6* (7), 544−556.

(56) Panovska-Griffiths, J.; Page, K. M.; Briscoe, J. A Gene Regulatory Motif That Generates Oscillatory or Multiway Switch Outputs. *J. R Soc. Interface* **2013**, *10* (79), 20120826.